Modelo de Regresión Múltiple

Profesor: Ricardo Pasquini

FCE and IAE, Universidad Austral

October 5, 2024

Objetivos de Aprendizaje

- Definir un modelo de regresión multiple.
- Comprender las diferencias de interpretacion de parametros respecto a la regresión simple.
- Comprender las diferencias en términos de propiedades estadísticas de los estimadores.
- Entender el problema clásico de "Inclusión de variable irrelevante".
- Entender el problema clásico de "Omisión de variable relevante".

Regresión Múltiple - Motivación

- Nos interesa agregar al modelo una cantidad arbitraria de variables explicativas. Por eso vamos a estudiar los modelos del tipo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

donde $x_1, x_2, ..., x_k$ representan k variables explicativas.

- Para la interpretación tomemos un ejemplo. Anteriormente estudiamos el modelo del ingreso (ing) de las personas en base a los años de educación ($educ$). Ahora incorporamos adicionalmente la experiencia (exp):

$$ing = \beta_0 + \beta_1 educ + \beta_2 exp + \epsilon$$

  - Interpretamos β_1: este coeficiente mide el efecto de un año de educación, una vez que *ya tenemos en cuenta* el efecto de los años de experiencia.

Comparación de Modelos

Modelo con Experiencia

$$ing = \beta_0 + \beta_1 educ + \beta_2 exp + \epsilon \tag{1}$$

- Interpretamos β_1: este coeficiente mide el efecto de un año de educación, una vez que *ya tenemos en cuenta* el efecto de los años de experiencia. También podemos interpretar como "manteniendo la experiencia constante".

Modelo sin Experiencia

$$ing = \beta_0 + \beta_1 educ + \epsilon \tag{2}$$

- En este caso, β_1 capturaba el efecto de un año extra de educación. Pero, ¿qué pasaría si la gente que se educa más siempre tiene más experiencia?
- Ya que un año extra de educación también implica mayor experiencia, entonces lo que capturaría β_1 sería el efecto del año extra de educación pero conteniendo el efecto de la experiencia.

Regresión Múltiple (cont.)

Estimador y Propiedades Estadísticas

- Las estimaciones del modelo de los parámetros (i.e., $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$) se obtienen mediante Mínimos Cuadrados Ordinarios (OLS).
- ¿Qué sabemos sobre las propiedades estadísticas de estos estimadores?
  1. Sabemos que estos estimadores son *insesgados*.
  2. Sabemos que la varianza toma la siguiente forma:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

Componentes de la Varianza

Fórmula de la Varianza de $\hat{\beta}_j$

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

donde:

- σ^2 es la varianza del error (n.b., que es desconocida pero estimable como ya vimos anteriormente).
- SST_j es la suma de los cuadrados totales de la j-ésima variable explicativa ($SST_j = \sum_{i=1}^{n}(x_{i,j} - \bar{x_j})^2$). La suma de los cuadrados totales es una medida de varianza, en este caso de la variable j-ésima.
- R_j^2 es una medida del tipo R^2 de bondad de ajuste que ya vimos. En este caso, el subíndice j indica que esta medida es el R^2 de otra regresión (no la que estamos analizando). R_j^2 se refiere a la bondad de ajuste del modelo que busca explicar x_j en función del resto de las variables explicativas del modelo.

Regresión Múltiple (cont.)

Inclusión de variables irrelevantes

- Incluir una variable irrelevante no implicará un sesgo en la estimación.
- La inclusión de una variable irrelevante implica una pérdida de eficiencia en la estimación.

Inclusión de Variables Irrelevantes

No hay sesgo en la estimación

- Supongamos que el modelo poblacional (el generador de los datos) es:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

- Pero al estimar el modelo, agregamos una variable x_2 como explicativa:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \epsilon$$

- En realidad, lo que se hizo no implica que se asumió un modelo incorrecto, ya que el modelo poblacional es equivalente al modelo estimado pero cuando el coeficiente de efecto de x_2 es 0:

$$y = \beta_0 + \beta_1 x_1 + 0 x_2 + \epsilon$$

- No hay inconsistencia entre el modelo poblacional y el estimado, y al estimar el modelo extendido, deberíamos esperar que el valor de $\hat{\beta}_2$ sea cercano a 0.

Regresión Múltiple (cont.)

Omision de variables relevantes

- La omisión de una variable relevante implica un sesgo en los coeficientes estimados.
- Se puede demostrar que la estimación estará sesgada.

Omisión de Variables Relevantes

Sesgo en los Coeficientes Estimados

- Supongamos que el modelo poblacional (el generador de los datos) es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- Pero al estimar el modelo, no contamos con x_2 como explicativa:

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \epsilon$$

- Se puede demostrar que la estimación estará sesgada.
  - Es relativamente fácil demostrar que:

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \delta_1$$

donde δ_1 es el coeficiente de la pendiente de una regresión de x_2 sobre x_1.

Omisión de Variables Relevantes

Sesgo en los Coeficientes Estimados

- El sesgo en la estimación de β_1 está determinado por la magnitud de $\beta_2\delta_1$.
- La magnitud del sesgo se anularía si:
  - x_2 no tiene efecto en y (i.e., en la población $\beta_2 = 0$), o bien,
  - x_2 no tiene relación con x_1. (i.e., $\delta_1 = 0$.)